Annotating language data

Tomaž Erjavec

Institut für Informationsverarbeitung
Geisteswissenschaftliche Fakultät
Karl-Franzens-Universität Graz

Lecture 3: Treebanks
17.11.2006

Overview

1. syntactic annotation and treebanks
   - lab work: TIGERSearch
2. lexical semantics
   - lab work: WordNet
3. Projects

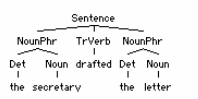
Treebanks

A linguistically annotated corpus that includes some grammatical analysis beyond word-level syntactic annotation (part-of-speech)

- "treebank" vs. "annotated corpus"
  - the first has to be manually annotated or post-edited
- two syntactic frameworks:
  - constituent structure
  - dependency structure

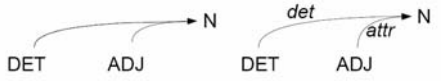
Constituent structure

- American structuralism, e.g. Zelig Harris (1951)
- "Bracketing": sentences consist of hierarchically embedded subparts → constituents
  - strings of words that belong together
  - constituency tests: substitution, movement, stand-alone test,…
- Part-whole relations
  - e.g. a NP consists of a determiner, adjective and noun
    $[_{NP}\ [_{DET}\ [_{ADJ}\ ]\ [_{N}\ ]]$

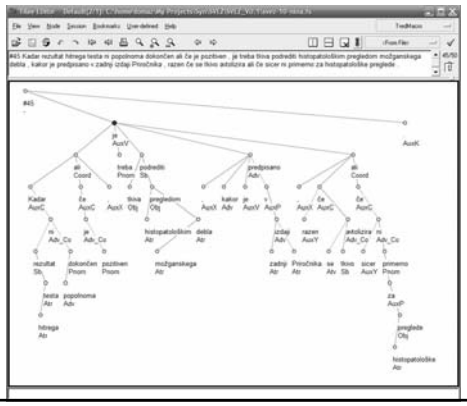


Dependency structure

- First comprehensive theory: Lucien Tesniere (1959)
- Sentence consists of hierarchically structured asymmetric binary relations between word forms → dependency relations (connections)
  - governor, dependent(s)
  - closely related to functional analysis
- Relations
  - e.g. determiner and adjective are subordinated to the noun



Dependencies in SDT



Hybrid models

Combine constituent and functional (dependency) information

- e.g. function added as additional sub-label to daughter category:
  [S [NP-SB …]] in Penn Treebank II

Treebanks and linguistic theory

- **Constituent structure,** e.g.
  - Penn Treebank I (AE)
- **Dependency structure, e.g.**
  - Prague Dependency Treebank / analytical level (Czech)
- **Constituent / Dependency Hybrid approaches**, e.g.
  - Penn Treebank II, SUSANNE (AE)
  - NEGRA/TIGER, TüBa (German)
- **Theory specific annotation,** e.g.
  - Prague Dependency Treebank / tectogrammatical level - Functional Generative Grammar
  - CCG-bank - Combinatory Categorial Grammar

Penn Treebank

- English treebank built at the University of Pennsylvania, distributed by LDC http://www.ldc.upenn.edu/
- Phase 1 (1989 - 1992)
  - skeletal parse
  - 2.6. mill words PoS tagged from Wall Street Journal, also other components, e.g. Brown Corpus
- Phase II (1993-1995)
  - enriching part of the material with grammatical functions and semantic relations
  - null-elements, coreference
- Phase III (1996-2000)
  - additional material: corpus of telephone conversations annotated for disfluencies

Penn Treebank: PoS annotation

- uses modified BROWN tagset
- allows multiple tags on word when annotator is unsure (avoid arbitrary decisions)
- 36 PoS tags, 12 other tags (punctuation, currency symbols)

1. CC Coordinating conj.	25.TO to
2. CD Cardinal number	26.UH Interjection
3. DT Determiner	27.VB Verb, base form
4. EX Existential there	28.VBD Verb, past tense
5. FW Foreign word	29.VBG V, gerund/pres. participle
6. IN Preposition/subord. conjunction	30.VBN Verb, past participle
7. JJ Adjective	31.VBP V, non-3rd ps.sing.present
8. JJR Adjective, comp.	32.VBZ V, 3rd ps.sing. present
9. JJS Adjective, superl.	33.WDT wh-determiner
10.LS List item marker	34.WP wh-pronoun
11.MD Modal	35.WP Possessive wh-pronoun
12.NN Noun, sg. or mass	36.WRB wh-adverb
13.NNS Noun, plural	37. # Pound sign
14.NNP Proper noun, singular	38. $ Dollar sign
15.NNPS Proper noun, plural	39. . Sentence-final punctuation
16.PDT Predeterminer	40. , Comma

etc.

Penn Treebank: syntactic annotation

1.	ADJP	Adjective phrase
2.	ADVP	Adverb phrase
3.	NP	Noun phrase
4.	PP	Prepositional phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by subordinating conjunction or 0 (zero 'that')
7.	SBARQ	Direct question introduced by wh-word or wh-phrase
8.	SINV	Declarative sentence with subject-aux inversion
9.	SQ	Subconstituent of SBARQ excluding wh-word or wh-phrase
10.	VP	Verb phrase
11.	WHADVP	Wh-adverb phrase
12.	WHNP	Wh-noun phrase
13.	WHPP	Wh-prepositional phrase
14.	X	Constituent of unknown or uncertain category

Penn Treebank: Skeletal parsing

```
( (S
   (NP Martin Marietta Corp.)
    was
    (VP given
        (NP a
            $ 29.9
            million Air Force contract
            (PP for
                (NP low-altitude navigation
                   and
                    targeting equipment)))))
 \
```

Penn Treebank: Functional tagset

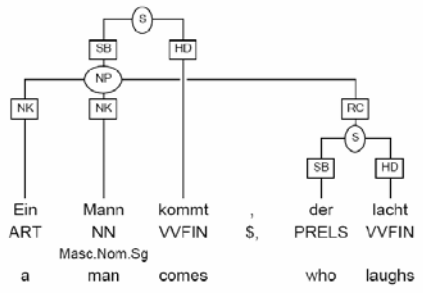
- Text categories
  - -HNL headlines and datelines
  - -LST list markers
  - -TTL titles
- Grammatical functions
  - -NOM non NPs that function as NPs
  - -ADV clausal and NP adverbials
  - -SBJ surface subject
  - …
- Semantic roles
  - -DIR direction and trajectory
  - -LOC location
  - -MCR manner
  - …
- Pseudo-attachment
  - *EXP* expletive
  - *RNR* right node raising
  - …

TIGER Treebank

"Linguis**TI**c Interpretation of a **GER**man Corpus"

- 50.000 sentences
- follow-up of NEGRA corpus (20.000 sentences)
- German newspaper texts (Frankfurter Rundshau)
- free licence
- hybrid annotation
- crossing branches for discontinuous constituents

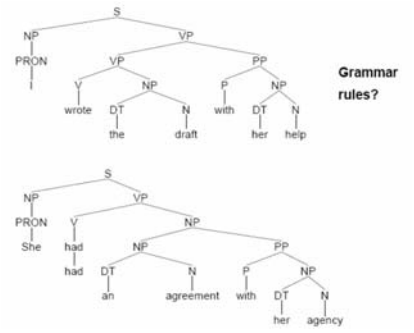
TIGER treebank example: discontinuous constituents



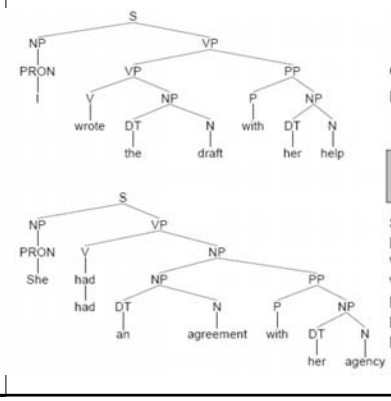
Creating treebanks

- Manual annotation
  - TrEd, CLaRK, Word freak
- Automatic annotation with human post-editing
  - Collins' Parser, Stanford Parser,…
- very labour intensive!

Exploiting treebanks: Parser training



Exploiting treebanks: Parser training



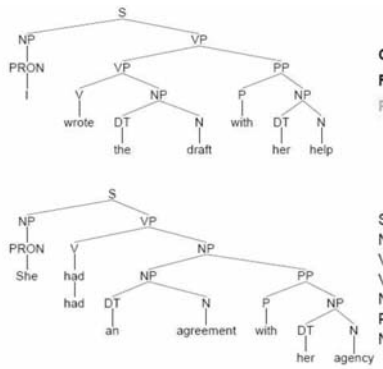
Grammar rules?

Frequencies?

Context-free grammar (CFG)

$S \rightarrow NP\ VP$
$NP \rightarrow PRON$
$VP \rightarrow VP\ PP$
$VP \rightarrow V\ NP$
$NP \rightarrow DT\ N$
$PP \rightarrow P\ NP$
$NP \rightarrow NP\ PP$

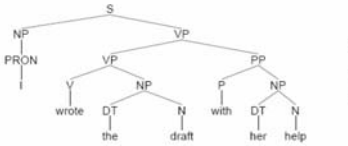
Exploiting treebanks: Parser training



Grammar rules?
Frequencies?
Probabilities?

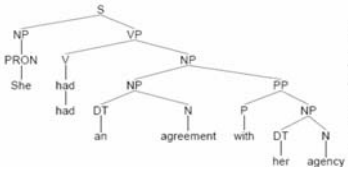
S → NP VP	2
NP → PRON	2
VP → VP PP	1
VP → V NP	2
NP → DT N	4
PP → P NP	2
NP → NP PP	1

Exploiting treebanks: Parser training



Grammar rules?
Frequencies?
Probabilities?

probabilistic context-free grammar (PCFG)



S → NP VP	2 / 2	1
NP → PRON	2 / 7	0.3
VP → VP PP	1 / 3	0.3
VP → V NP	2 / 3	0.7
NP → DT N	4 / 7	0.6
PP → P NP	2 / 2	1
NP → NP PP	1 / 7	0.1

Exploiting treebanks: Charniak 1996

- inducing a treebank-based PCFG
- preliminary version of Penn Treebank
- training corpus: ~30,000 words
- test corpus: ~30,000 words

Sentence Length	Average Length	Precision	Recall
2-12	8.7	88.6	91.7
2-16	11.4	85.0	87.7
2-20	13.8	83.5	86.2
2-25	16.3	82.0	84.0
2-30	18.7	80.6	82.5
2-40	21.9.	78.8	80.4

CoNLL-X shared task on multilingual dependency parsing

- 2006, http://nextens.uvt.nl/~conll/
- open task: common format of treebanks, all systems must compete on all languages
- 13 treebanks: Arabic, Chinese, Czech, Danish, Dutch, German, Japanese, Portuguese, Slovene, Spanish, Swedish, Turkish, Bulgarian
- 20 systems
- Best average labelled attachment score ~80%